BRIEF COMMUNICATION

Validation of a Model of Lung Cancer Risk Prediction Among Smokers

Kathleen A. Cronin, Mitchell H. Gail, Zhaohui Zou, Peter B. Bach, Jarmo Virtamo, Demetrius Albanes

The Bach model was developed to predict the absolute 10-year risk of developing lung cancer among smokers by use of participants in the Carotene and Retinol Efficacy Trial of lung cancer prevention. We assessed the validity of the Bach model among 6239 smokers from the placebo arm of the Alpha-Tocopherol, Beta-Carotene Cancer Prevention (ATBC) Study. The expected numbers of lung cancer cases and deaths without lung cancer were calculated from the Bach model and compared with the observed numbers of corresponding events over 10 years. We found that the risk model slightly underestimated the observed lung cancer risk (number of lung cancers expected/number observed = 0.89, 95% confidence interval [CI] = 0.80 to 0.99) over 10 years. The competing risk portion of the model substantially underestimated risk of non-lung cancer mortality (number of non-lung cancer deaths expected/number observed = 0.61, 95% CI = 0.57 to 0.64) over 10 years. The age-specific concordance indices for 10-year predictions were 0.77 (95% CI = 0.70 to 0.84), 0.59 (95% CI = 0.53 to 0.65), 0.62 (95% CI = 0.57 to 0.67), and 0.57 (95% CI = 0.49 to 0.67) for the age groups 50–54, 55–59, 60–64, and 65–69 years, respectively. Periodic radiographic screening in the ATBC Study may explain why slightly more cancers were observed than expected from the Bach model. [J Natl Cancer Inst 2006;98:637–40]

Bach et al. *(1)* used data from the Carotene and Retinol Efficacy Trial (CARET) *(2,3)* to develop a model that predicts the probability or absolute risk of being diagnosed with lung cancer—i.e., lung cancer risk measured in the presence of competing causes of death over a 10-year period that is based on an individual's age, sex, asbestos exposure history, and smoking history (duration, number of cigarettes per day, and time since quitting smoking). Such an absolute risk model is useful for counseling smokers and for designing intervention trials, because the power of such trials depends on the number of incident lung cancers, which is proportional to the average absolute risk. To compute absolute risk, Bach et al. developed models for the pure risk of lung cancer and for competing risks of mortality from other causes. The absolute risk calculation can be inaccurate if either the model for pure lung cancer risk or the model for competing risks is inaccurate. Using independent data for 6239 smokers from the placebo arm of the Alpha-Tocopherol, Beta-Carotene Cancer Prevention (ATBC) Study *(4)* that met the recommended conditions for the risk model, we assessed the validity of the 10-year absolute risk predictions from the CARET model, as well as the pure lung cancer risk and competing risk components.

The ATBC Study data provide a stringent test of the Bach model because of differences in the populations and surveillance regimes. The ATBC Study included men aged 50–69 years from southwestern Finland who smoked five or more cigarettes per day at entry, whereas CARET recruited males and females who were heavy smokers aged 50–69 years with at least 20 pack-years of smoking exposure including recent quitters, and it also recruited asbestos-exposed men aged 45–69 years who were current or former smokers from multiple centers in the United States. The ATBC Study used the current number of cigarettes smoked per day to measure smoking intensity, whereas CARET used the average number of cigarettes smoked per day when smoking to measure smoking intensity. Subjects in CARET had a chest x-ray before entering the trial *(2)*, whereas those in the ATBC Study had both screening chest x-rays at study entry and periodic screening examinations averaging every 2 years and 4 months during the study (i.e., every seventh clinic visit) and at the end of the study *(4)*.

The supplemental equations in Bach et al. *(1)* (available at http://jncicancerspectrum.oxfordjournals.org/jnci/content/vol95/issue9) defined two proportional hazards regression models to estimate the pure probability of developing lung cancer in 1 year and the competing risk of dying without lung cancer in 1 year. To predict the absolute lung cancer risk through T years of follow-up, the two single-year models were applied recursively T times *(1)*. To evaluate the T-year absolute risk predictions, we assumed that smoking patterns at baseline persisted (i.e., current smokers were assumed to continue smoking), and accordingly, we increased the age and duration of smoking by one for each year of follow-up. For a T equal to 10 years, each individual contributed 10 years to the predicted risk, even if he was diagnosed with lung cancer or died during follow-up. The sum of the individuals' predicted risks was the expected number of lung cancer cases over the 10 years, and this quantity was compared with the number of observed cases, as described below. Similar methods were used to compute the observed and expected numbers of deaths from non-lung cancer causes.

To evaluate the models for pure lung cancer risk (lung cancer risk in the absence of competing causes of death) and competing risks (risk of death in the absence of lung cancer), we focused on

Affiliations of authors: Division of Cancer Control and Population Sciences (KAC), Division of Cancer Epidemiology and Genetics (MHG, DA), National Cancer Institute, Bethesda, MD; Information Management Services, Inc., Silver Spring, MD (ZZ); Memorial Sloan-Kettering Cancer Center, New York, NY (PBB); Department of Epidemiology and Health Promotion, National Public Health Institute, Helsinki, Finland (JV).

Correspondence to: Kathleen A. Cronin, PhD, Statistical Research and Applications Branch, 6116 Executive Boulevard, Suite 504, Bethesda, MD 20892–8317 (e-mail: cronink@mail.nih.gov).

See "Notes" following "References."

DOI: 10.1093/jnci/djj163

Table 1. Ten-year absolute risk prediction for lung cancer*

	Lung cancer			Death without lung cancer		
	No. of lung cancer events			No. of deaths		
Variable	*O*	*E*	*E*/*O* ratio (95% CI)	*O*	*E*	*E*/*O* ratio (95% CI)
Overall	333	297.07	0.89 (0.80 to 0.99)	1139	691.92	0.61 (0.57 to 0.64)
Age group						
50–54 y	53	62.91	1.19 (0.91 to 1.55)	257	121.64	0.47 (0.42 to 0.53)
55–59 y	99	94.93	0.96 (0.79 to 1.17)	341	190.51	0.56 (0.50 to 0.62)
60–64 y	136	89.97	0.66 (0.56 to 0.78)	316	219.16	0.69 (0.62 to 0.77)
65–69 y	45	48.86	1.09 (0.81 to 1.45)	223	159.18	0.71 (0.63 to 0.81)
Cigarettes per day, No.						
<20	96	68.48	0.71 (0.58 to 0.87)	382	222.15	0.58 (0.53 to 0.64)
20–29	175	156.69	0.90 (0.77 to 1.04)	548	340.58	0.62 (0.57 to 0.68)
≥30	62	71.90	1.16 (0.90 to 1.49)	209	129.19	0.62 (0.54 to 0.71)
Duration of smoking						
<35 y	32	46.97	1.47 (1.04 to 2.08)	251	122.48	0.49 (0.43 to 0.55)
35–39 y	68	64.01	0.94 (0.74 to 1.19)	227	129.46	0.57 (0.50 to 0.65)
40–44 y	126	103.19	0.82 (0.69 to 0.98)	369	224.35	0.61 (0.55 to 0.67)
45–49 y	85	62.89	0.74 (0.60 to 0.92)	214	159.22	0.74 (0.65 to 0.85)
50–55 y	22	20.20	0.91 (0.60 to 1.39)	78	56.41	0.72 (0.58 to 0.90)

**O* = observed; *E* = expected; CI = confidence interval; y = years.

1-year predictions, within which the effects of competing risks were negligible. Follow-up covariate information from the ATBC Study was used to estimate single-year risk for 10 years. For example, if a person quit smoking after 5 years of follow-up, his risk would be predicted as a smoker in the first 5 years and as a former smoker with increasing years of abstinence in the subsequent 5 years. In this analysis, a person must survive event-free to continue to contribute to the expected number of lung cancers in successive years. To conform to the recommended ranges of model predictors *(1)*, we excluded individuals aged 75 years or older and individuals with smoking durations of more than 55 years. By focusing on single-year predictions and hence on pure lung cancer and competing risks, we obtained insight into how these two components explain discrepancies between observed and expected absolute risks over longer follow-up periods.

For both absolute and pure risk analyses, we compared the expected number (*E*) of lung cancer cases and deaths without lung cancer to the observed numbers (*O*) in the ATBC Study control population by determining the *E*/*O* ratio. The 95% confidence interval (CI) for this ratio was calculated as follows:

$$\frac{E}{O}\exp\left(\pm 1.96\times\frac{1}{\sqrt{O}}\right).$$

Although most comparisons of expected numbers with observed numbers set *T* equal to 10 for absolute risk (Table 1) and *T* equal to 1 for pure risks (Table 2),

Table 2. One-year risk prediction for lung cancer*

	Lung cancer			Death without lung cancer		
	No. of lung cancer events			No. of deaths		
	O	*E*	*E*/*O* ratio (95% CI)	*O*	*E*	*E*/*O* ratio (95% CI)
Overall	321	258.50	0.81 (0.72 to 0.90)	1068	590.26	0.55 (0.52 to 0.59)
Duration of follow-up						
1–2 y	35	40.85	1.17 (0.84 to 1.63)	152	87.56	0.58 (0.49 to 0.68)
3–4 y	65	47.98	0.74 (0.58 to 0.94)	204	104.97	0.51 (0.45 to 0.59)
5–6 y	83	53.80	0.65 (0.52 to 0.80)	217	122.40	0.56 (0.49 to 0.64)
7–8 y	72	56.75	0.79 (0.63 to 0.99)	233	132.98	0.57 (0.50 to 0.65)
9–10 y	66	59.12	0.90 (0.70 to 1.14)	262	142.34	0.54 (0.48 to 0.61)
Age group						
50–54 y	3	10.27	3.42 (1.10 to 10.62)	59	22.81	0.39 (0.30 to 0.50)
55–59 y	52	49.49	0.95 (0.73 to 1.25)	201	96.12	0.48 (0.42 to 0.55)
60–64 y	98	86.89	0.89 (0.73 to 1.08)	330	172.43	0.52 (0.47 to 0.58)
65–69 y	124	78.85	0.64 (0.53 to 0.76)	293	191.20	0.65 (0.58 to 0.73)
70–74 y	44	32.99	0.75 (0.56 to 1.01)	185	107.71	0.58 (0.50 to 0.67)
Cigarettes per day, No.						
<20	93	57.59	0.62 (0.51 to 0.76)	355	181.14	0.51 (0.46 to 0.57)
20–29	169	136.85	0.81 (0.70 to 0.94)	513	293.24	0.57 (0.52 to 0.62)
≥30	59	65.19	1.10 (0.86 to 1.43)	200	115.87	0.58 (0.50 to 0.67)
Duration of smoking						
<35 y	9	13.38	1.49 (0.77 to 2.86)	111	48.57	0.44 (0.36 to 0.53)
35–39 y	31	38.54	1.24 (0.87 to 1.77)	184	91.19	0.50 (0.43 to 0.57)
40–44 y	90	72.70	0.81 (0.66 to 0.99)	257	147.63	0.57 (0.51 to 0.65)
45–49 y	117	81.94	0.70 (0.58 to 0.84)	315	175.42	0.56 (0.50 to 0.62)
50–55 y	74	51.93	0.70 (0.56 to 0.88)	201	127.46	0.63 (0.55 to 0.73)

**O* = observed; *E* = expected; CI = confidence interval; y = years.

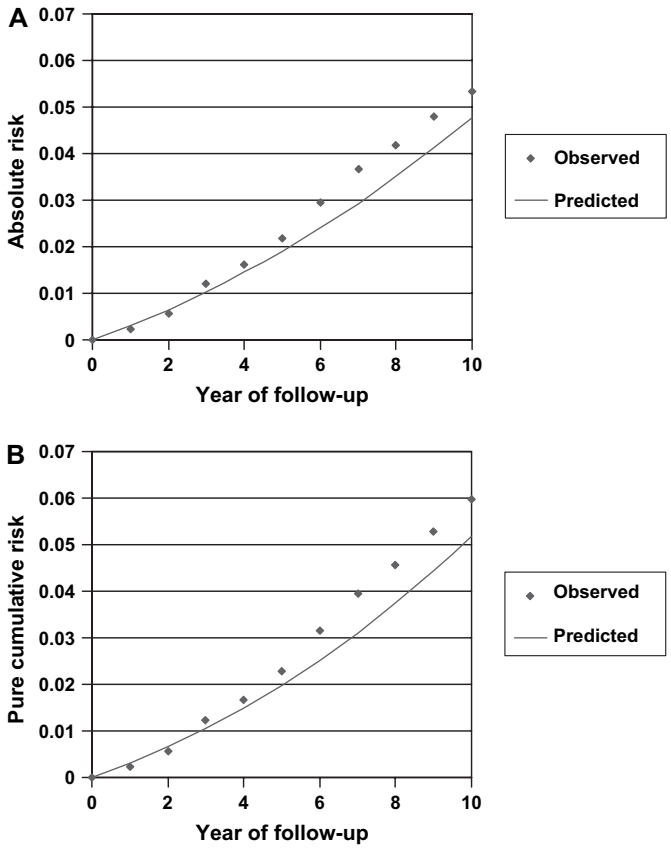

Fig. 1. Predicted risk and observed risk from the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study plotted by year of follow-up. **A**) Observed proportion of the participants in the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study with incident lung cancer and the corresponding average predicted absolute lung cancer risk from the Bach model. Both values are plotted against years of follow-up. **B**) Observed cumulative pure lung cancer risk in participants of the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study and cumulative pure lung cancer risk predicted from the Bach model. Both values are plotted against years of follow-up. **Solid diamonds** = observed; **solid line** = predicted.

we also graphed the comparisons for various follow-up times. For absolute lung cancer risk, we plotted the observed proportion of the population with incident lung cancers against follow-up time, and for comparison, we plotted the average expected absolute risk from the Bach model (Fig. 1, A). To evaluate the model for pure lung cancer risk, we plotted one minus the Kaplan–Meier estimate of lung cancer risk for the entire population against follow-up time; in this calculation, deaths were treated as independent censoring events. For comparison, we plotted the average pure cumulative lung cancer risk against follow-up time (Fig. 1, B). For each individual, the pure cumulative risk through year T was calculated as $[1 - \prod_{1}^{T} (1 - \text{single-year pure lung cancer risk})]$.

The number of observed lung cancers at 10 years slightly exceeded the absolute risk model predictions (overall E/O ratio = 0.89, 95% CI = 0.80 to 0.99) (Table 1). The underestimation of risk is more pronounced for those who smoked fewer than 20 cigarettes per day (Tables 1 and 2). Although the model gives reasonably accurate 10-year estimates of lung cancer incidence, it substantially underestimates the 10-year risk of non-lung cancer mortality (overall E/O ratio = 0.61, 95% CI = 0.57 to 0.64) (Table 1). There were 64% (i.e., $100 \times [1/0.61 - 1]$) more deaths than predicted.

The overall E/O ratio for pure lung cancer risk, approximated by 1-year predictions (Table 2), was 0.81 (95% CI = 0.72 to 0.90). For absolute lung cancer risk calculations over longer time intervals, this underestimation in pure lung cancer hazard was partly offset by the underestimation of the hazard of competing risks of mortality, which explains why the E/O ratio of 0.89 for absolute lung cancer risk in Table 1 is closer to unity than the corresponding ratio in Table 2.

The discrepancy between the observed proportion of the population with lung cancer and the expected proportion obtained from the Bach absolute risk model increases with duration of follow-up (Fig. 1, A). Agreement is quite good through 5 years but degrades thereafter. A similar pattern was observed for pure lung cancer risk (Fig. 1, B), although the discrepancies between the empirical and model-based estimates were somewhat greater than in Fig. 1, A. The overall concordance index *(5)*, which estimates the probability that a subject who develops lung cancer will have a higher predicted 10-year risk than a subject who does not develop lung cancer, was equal to 0.69 (95% CI = 0.66 to 0.72) and was similar to the 0.72 concordance index reported in the original paper *(1)*. To evaluate the discriminatory power of covariates other than age, we computed the concordance index within 5-year age groups. The results were 0.77 (95% CI = 0.70 to 0.84), 0.59 (95% CI = 0.53 to 0.65), 0.62 (95% CI = 0.57 to 0.67), and 0.57 (95% CI = 0.49 to 0.67) for age groups of 50–54, 55–59, 60–64, and 65–69 years, respectively.

Two criteria are often used to assess the validity of risk projection models: calibration, which is based on a comparison of observed with expected number, and discriminatory power, which is measured by the concordance statistic. For calibration, the Bach model underestimated the absolute lung cancer risk in the ATBC Study by 11% overall (Table 1), which is satisfactory for many purposes. Although the pure lung cancer risks were underestimated by 19% overall (Table 2), this feature was partially compensated for in computing absolute lung cancer risk by offsetting underestimates in the competing hazard of mortality (Table 2).

The higher observed lung cancer incidence in the ATBC Study, compared with that expected from the Bach model, probably reflects surveillance with periodic chest x-rays in the former. Surveillance with chest x-rays in a screened group resulted in higher lung cancer incidence than in a nonscreened control group both in the Mayo Lung Project and in a randomized trial of chest x-ray screening conducted in Czechoslovakia *(6,7)*. Data in Table 2 support this screening hypothesis, because the E/O ratio was greater than 1.0 in follow-up years 1–2, shortly after both study populations had received chest x-rays, and was approximately 1.0 in years 9–10, a period in which participants in the ATBC Study no longer received study-based x-rays. These data highlight

the importance of surveillance patterns in determining absolute risk. Another difference that may have influenced our results was that the ATBC Study used the current number of cigarettes smoked per day, whereas CARET used the average number smoked per day when smoking. The direction of the bias introduced by this difference is uncertain.

The discriminatory power of the Bach model, measured by age-specific concordance indices, was comparable to the range of 0.58–0.63 reported for breast cancer risk models *(8–10)*. That these statistics are not closer to 1.0 reflects the general challenge of predicting cancer risk, even in cases where the risk factors are well known and measurable.

References

(1) Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ, et al. Variations in lung cancer risk among smokers. J Natl Cancer Inst 2003;95:470–8.

(2) Omenn GS, Goodman PC, Thornquist M, Grizzle J, Rosenstock L, Barnhart S, et al. The beta-carotene and retinol efficacy trial (CARET) for chemoprevention of lung cancer in high risk populations: smokers and asbestos-exposed workers. Cancer Res 1994; 54(7 Suppl):2038s–43s.

(3) Omenn GS, Goodman GE, Thornquist MD, Balmes J, Cullen MR, Glass A, et al. Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. New Engl J Med 1996;334:1150–5.

(4) Albanes D, Heinonen OP, Taylor PR, Virtamo J, Edwards BK, Rautalahti M, et al. α-Tocopherol and β-carotene supplements and lung cancer incidence in the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study: effects of base-line characteristics and study compliance. J Natl Cancer Inst 1996;88:1560–70.

(5) DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparameteric approach. Biometrics 1988;44:837–45.

(6) Fontana, RS, Sanderson DR, Woolner LB, Taylor WF, Miller WE, Muhm JR, et al. Screening for lung cancer: a critique of the Mayo Lung Project. Cancer 1991;67:1155–64.

(7) Rockhill B, Spiegelman D, Byrne C, Hunter DJ, Colditz GA. Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. J Natl Cancer Inst 2001;93:358–66.

(8) Bach PB, Kelley MJ, Tate RC, McCrory DC. Screening for lung cancer: a review of the current literature. Chest 2003;123: 72s–82s.

(9) Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. Biostatistics 2005;6:227–39.

(10) Rockhill B, Byrne C, Rosner B, Louie MM, Colditz G. Breast cancer risk prediction with a log-incidence model: evaluation of accuracy. J Clin Epidemiol 2003;56:856–61.

Notes

Funding to pay the Open Access publication charges for this article was provided by the Division of Cancer Control and Population Sciences, National Cancer Institute.

Manuscript received July 11, 2005; revised February 9, 2006; accepted February 17, 2006.